Efficient Maximum Likelihood Estimation of a Low-Rank Probability Mass Tensor from Partial Observations

Martin Haardt

Ilmenau University of Technology
Communications Research Laboratory
98684 Ilmenau, Germany

E-Mail: Martin.Haardt@tu-ilmenau.de

Homepage: http://www.tu-ilmenau.de/crl

Acknowledgments

- Joint work with
  - ⇒ Arie Yeredor, Tel Aviv University, Israel
  - ⇒ Joseph Chege, TU Ilmenau, Germany

Outline

- Motivation for Estimating the joint PMF Tensor of a Set of Random Variables
- Modeling of the Naïve Bayes Model via the CPD
- Joint PMF Tensor Estimation Methods
  - ⇒ Marginal-based Estimation
    - LS (Least Squares) Criterion
    - KLD (Kullback-Leibler Divergence) Criterion
  - ⇒ Data-based Estimation
    - AD (Alternating Directions) Minimization
    - EM (Expectation Maximization)
- Simulation Results
  - ⇒ Synthetic Data Experiments
    - Speeding up the Convergence of EM: SQUAREM
  - ⇒ Real Data Experiments
- Conclusions

Outline

- Motivation for Estimating the joint PMF Tensor of a Set of Random Variables
- Modeling of the Naïve Bayes Model via the CPD
- Joint PMF Tensor Estimation Methods
  - ⇒ Marginal-based Estimation
    - LS (Least Squares) Criterion
    - KLD (Kullback-Leibler Divergence) Criterion
  - ⇒ Data-based Estimation
    - AD (Alternating Directions) Minimization
    - EM (Expectation Maximization)
- Simulation Results
  - ⇒ Synthetic Data Experiments
    - Speeding up the Convergence of EM: SQUAREM
  - ⇒ Real Data Experiments
- Conclusions

Motivation (1)

- Estimating the **p**robability **m**ass **f**unction (**PMF**) from partial observations of a discrete random vector is a core problem in statistical learning
- Knowledge of the joint PMF is considered as the gold-standard
  - ⇒ enables optimal estimation under criteria such as
    - **m**ean **s**quared **e**rror (**MSE**) and
    - minimum probability of error
- Useful applications include
  - ⇒ **recommender systems** – using the joint PMF of ratings,
    - predict missing ratings based on the conditional expectation of the available ratings and
    - make recommendations based on these predictions
  - ⇒ **data classification** – using the joint PMF of the features and the label,
    - infer the label of a test sample according to the MAP principle

Motivation (2)

- Usually, the full PMF is unknown
  - ⇒ needs to be estimated from the data during a ‘training’ stage
- Estimating the joint PMF via histograms considered virtually impossible
  - ⇒ curse of dimensionality, e.g., 10 variables with 10 distinct values each = 10^{10} parameters to be estimated!
  - ⇒ histogram estimation very inaccurate – low probability of encountering any particular realization
  - ⇒ a lot of data is required to obtain a reliable estimate (even more in the case of incomplete data!)

Motivation (3)

- Consider a set of N random variables $\{X_n\}_{n=1}^{N}$
- Each variable X_n takes I_n discrete values
- Our aim is to estimate the joint PMF $\Pr(X_1, \ldots, X_N)$ from (partial) realizations of the random variables
- How would we use $\Pr(X_1, \ldots, X_N)$?
  - ⇒ data classification – given a set of features $X_1, \ldots, X_{N-1}$, compute the posterior probability and predict the label X_N using the MAP principle

$$\Pr(X_N = i_N | X_1 = i_1, \ldots, X_{N-1} = i_{N-1}) = \frac{\Pr(X_1 = i_1, \ldots, X_N = i_N)}{\sum_{i_N=1}^{I_N} \Pr(X_1 = i_1, \ldots, X_N = i_N)} \quad (1)$$

  - ⇒ recommender systems – given a subset of ratings $X_1, \ldots, X_{N-1}$, estimate X_N using the conditional expectation

$$\mathbb{E}\{X_N | X_1 = i_1, \ldots, X_{N-1} = i_{N-1})\} = \sum_{i_N=1}^{I_N} i_N \Pr(X_N = i_N | X_1 = i_1, \ldots, X_{N-1} = i_{N-1}) \quad (2)$$

Motivating Example: Recommendation of Movies (1)

- Consider 4 movies each with 5 possible ratings:
  - ⇒ Movie 1: Toy Story
  - ⇒ Movie 2: Shrek
  - ⇒ Movie 3: Inception
  - ⇒ Movie 4: The Batman
- The joint PMF of a user's ratings can be represented as a $5 \times 5 \times 5 \times 5$ **tensor** $\boldsymbol{\mathcal{X}}$
- For example, $\boldsymbol{\mathcal{X}}(1, 3, 4, 5)$ is the probability that the user rated Toy Story as 1, Shrek as 3, Inception as 4, and The Batman as 5

Motivating Example: Recommendation of Movies (2)

- Suppose this user has rated Toy Story (t), Inception (n), and The Batman (b), but has not watched Shrek
- If we know the joint PMF, we can easily predict the user's rating for Shrek
  - ⇒ calculate the **posterior probability** vector $\boldsymbol{r} \in \mathbb{R}^5$

$$\boldsymbol{r}(i) = \mathrm{Pr}(X_2 = i | X_1 = t, X_3 = n, X_4 = b) = \frac{\boldsymbol{\mathcal{X}}(t, i, n, b)}{\sum_{j=1}^{5} \boldsymbol{\mathcal{X}}(t, j, n, b)}$$

  - ⇒ calculate the **MMSE estimate** of the rating for Shrek

$$\hat{r} = \sum_{i=1}^{5} i \cdot \boldsymbol{r}(i)$$

  - ⇒ based on the estimated rating, decide whether or not to recommend Shrek to the user

Outline

CPD of a Joint PMF Tensor (1)

- The joint PMF $\text{Pr}(X_1, \ldots, X_N)$ can be represented as an N-way tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$
- Each element of $\boldsymbol{\mathcal{X}}$ is the probability of a particular realization / index combination, i.e.,

$$\boldsymbol{\mathcal{X}}(i_1, \ldots, i_N) = \text{Pr}(X_1 = i_1, \ldots, X_N = i_N) \tag{3}$$

- $\boldsymbol{\mathcal{X}}$ admits a rank-R **c**anonical **p**olyadic **d**ecomposition (**CPD**)

$$\boldsymbol{\mathcal{X}} = \sum_{r=1}^{R} \boldsymbol{\lambda}(r) \boldsymbol{A}_1(:, r) \circ \boldsymbol{A}_2(:, r) \circ \cdots \circ \boldsymbol{A}_N(:, r) \tag{4}$$

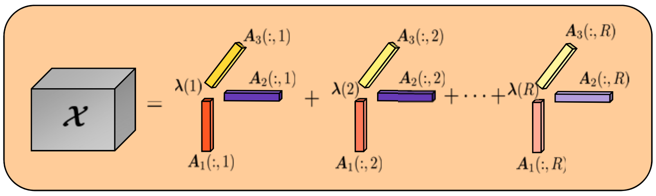
$$\triangleq [\![\boldsymbol{\lambda}; \boldsymbol{A}_1, \ldots, \boldsymbol{A}_N]\!]$$

$\boldsymbol{\lambda} \in \mathbb{R}^R \rightarrow$ loading vector
$\boldsymbol{A}_n \in \mathbb{R}^{I_n \times R} \rightarrow$ factor matrix
$\circ \rightarrow$ outer product

CPD of a Joint PMF Tensor (2)

- The probability of a particular index combination is given by

$$\begin{aligned}
\text{Pr}(X_1 = i_1, \ldots, X_N = i_N) &= \boldsymbol{\mathcal{X}}(i_1, \ldots, i_N) \\
&= \sum_{r=1}^{R} \boldsymbol{\lambda}(r) \boldsymbol{A}_1(i_1, r) \boldsymbol{A}_2(i_2, r) \cdots \boldsymbol{A}_N(i_N, r) \\
&= \sum_{r=1}^{R} \boldsymbol{\lambda}(r) \prod_{n=1}^{N} \boldsymbol{A}_n(i_n, r) \qquad (5)
\end{aligned}$$

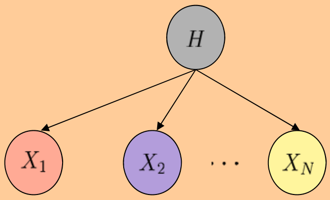


Naïve Bayes Model (1)

- In practice, random variables usually have some degree of dependence (i.e., not fully independent or fully dependent)

- A reasonable assumption: the random variables are independent given another hidden (latent) random variable H taking a finite set of R values

  $\Rightarrow$ also known as the naïve Bayes model

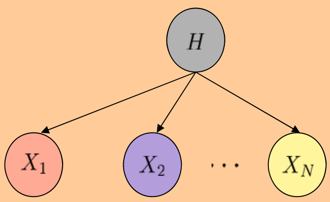


$X_1, \ldots, X_N$ are independent given the latent variable H

Naïve Bayes Model (2)

- The probability of a particular realization according to the model is

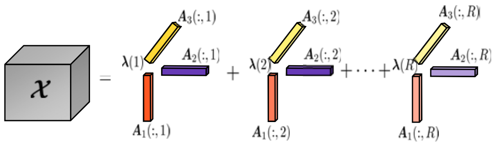
$$\begin{aligned}\mathsf{Pr}(X_1 = i_1, \ldots, X_N = i_N) &= \sum_{r=1}^{R} \mathsf{Pr}(H = r)\mathsf{Pr}(X_1 = i_1|H = r)\cdots\mathsf{Pr}(X_N = i_N|H = r)\\ &= \sum_{r=1}^{R} \mathsf{Pr}(H = r)\prod_{n=1}^{N} \mathsf{Pr}(X_n = i_n|H = r) \qquad (6)\end{aligned}$$



$X_1, \ldots, X_N$ are independent given the latent variable H

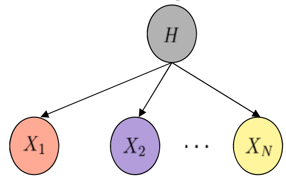
The Naïve Bayes Model and the CPD (1)

CPD



$$\Pr(X_1 = i_1, \ldots, X_N = i_N) = \sum_{r=1}^{R} \boldsymbol{\lambda}(r) \prod_{n=1}^{N} \boldsymbol{A}_n(i_n, r)$$

Naïve Bayes Model



$$\Pr(X_1 = i_1, \ldots, X_N = i_N) = \sum_{r=1}^{R} \Pr(H = r) \prod_{n=1}^{N} \Pr(X_n = i_n | H = r)$$

a priori probability $\boldsymbol{\lambda}(r) = \Pr(H = r)$

conditional (*a posteriori*) probability $\boldsymbol{A}_n(i_n, r) = \Pr(X_n = i_n | H = r), \; i_n \in [1, I_n]$

$\boldsymbol{\lambda} > \mathbf{0}, \; \mathbf{1}^{\mathsf{T}}\boldsymbol{\lambda} = 1$

$\boldsymbol{A}_n \geq \mathbf{0}, \; \mathbf{1}^{\mathsf{T}}\boldsymbol{A}_n = \mathbf{1}^{\mathsf{T}}, \; n = 1, \ldots, N$

Probability simplex constraints

$\Longrightarrow \boldsymbol{\mathcal{X}}$ admits a non-negative CPD

[1] N. Kargas, N. D. Sidiropoulos, and X. Fu, "Tensors, Learning, and "Kolmogorov Extension" for Finite-Alphabet Random Vectors," *IEEE Trans. Signal Process.*, vol. 66, no. 18, pp. 4854–4868, Sep. 2018.

The Naïve Bayes Model and the CPD (2)

- The naïve Bayes model leads to a nice Bayesian interpretation of the CPD
  - $\Rightarrow$ consider our earlier example with Toy Story, Shrek, Inception, and The Batman, with 5 possible ratings
  - $\Rightarrow$ if there are $R = 2$ kinds of users, i.e., action lovers and animation lovers, then we may have

$$\boldsymbol{\lambda} = [0.4, 0.6]^{\mathsf{T}}$$

	$\boldsymbol{A}_1$ Toy Story		$\boldsymbol{A}_2$ Shrek		$\boldsymbol{A}_3$ Inception		$\boldsymbol{A}_4$ The Batman	
1	0.20	0.05	0.50	0.04	0.02	0.20	0.04	0.45
2	0.40	0.10	0.20	0.01	0.08	0.40	0.06	0.35
3	0.30	0.25	0.15	0.15	0.15	0.30	0.15	0.10
4	0.05	0.35	0.10	0.45	0.20	0.05	0.40	0.05
5	0.05	0.25	0.05	0.35	0.55	0.05	0.35	0.05

The Naïve Bayes Model and the CPD (3)

- Implication: *any* joint PMF admits a naïve Bayes representation with a finite latent alphabet [1]
- The number of states R the latent variable H can take is bounded by

$$R \leq \min_k \left(\prod_{\substack{n=1 \\ n \neq k}}^{N} I_n \right) \tag{7}$$

- Uniqueness not guaranteed – in addition, CPD uniqueness conditions must be fulfilled

[1] N. Kargas, N. D. Sidiropoulos, and X. Fu, “Tensors, Learning, and “Kolmogorov Extension” for Finite-Alphabet Random Vectors,” *IEEE Trans. Signal Process.*, vol. 66, no. 18, pp. 4854–4868, Sep. 2018.

PMF Estimation Methods

- Two main approaches for the estimation of the PMF tensor
  - ⇒ marginal-based – estimate lower-order marginals and fit them to a CPD [1], [2]
  - ⇒ data-based – find maximum likelihood estimate of joint PMF using observed data [3]

[1] N. Kargas, N. D. Sidiropoulos, and X. Fu, “Tensors, Learning, and “Kolmogorov Extension” for Finite-Alphabet Random Vectors,” *IEEE Trans. Signal Process.*, vol. 66, no. 18, pp. 4854–4868, Sep. 2018.

[2] A. Yeredor and M. Haardt, “Estimation of a Low-Rank Probability-Tensor from Sample Sub-Tensors via Joint Factorization Minimizing the Kullback-Leibler Divergence,” in *Proc. 27th European Signal Processing Conference (EUSIPCO)*, A Coruna, Spain, Sep. 2019.

[3] A. Yeredor and M. Haardt, “Maximum Likelihood Estimation of a Low-Rank Probability Mass Tensor From Partial Observations,” *IEEE Signal Process. Lett.*, vol. 26, no. 10, pp. 1551–1555, Oct. 2019.

Outline

- Motivation for Estimating the joint PMF Tensor of a Set of Random Variables
- Modeling of the Naïve Bayes Model via the CPD
- Joint PMF Tensor Estimation Methods
  - ⇒ Marginal-based Estimation
    - LS (Least Squares) Criterion
    - KLD (Kullback-Leibler Divergence) Criterion
  - ⇒ Data-based Estimation
    - AD (Alternating Directions) Minimization
    - EM (Expectation Maximization)
- Simulation Results
  - ⇒ Synthetic Data Experiments
    - Speeding up the Convergence of EM: SQUAREM
  - ⇒ Real Data Experiments
- Conclusions

Marginal-Based Estimation

- The full joint PMF can be recovered from marginals of order ≥ 3 under relatively mild conditions [1]
- **Idea**: easier to obtain reliable empirical estimates of lower-order marginals
- Lower order marginal distributions also admit a non-negative CPD, e.g.,

$$
\begin{aligned}
\Pr(i_1,\ldots,i_{k-1},i_{k+1},\ldots,i_N) &= \sum_{i_k=1}^{I_k} \Pr(i_1,\ldots,i_N) \\
&= \sum_{r=1}^{R}\sum_{i_k=1}^{I_k} \boldsymbol{\lambda}(r) \prod_{n=1}^{N} \boldsymbol{A}_n(i_n,r) \\
&= \sum_{r=1}^{R} \boldsymbol{\lambda}(r) \prod_{\substack{n=1\\ n\neq k}}^{N} \boldsymbol{A}_n(i_n,r) \underbrace{\sum_{i_k=1}^{I_k} \boldsymbol{A}_k(i_k,r)}_{=1 \text{ due to probability simplex constraint}} \\
&= \sum_{r=1}^{R} \boldsymbol{\lambda}(r) \prod_{\substack{n=1\\ n\neq k}}^{N} \boldsymbol{A}_n(i_n,r) \qquad (8)
\end{aligned}
$$

[1] N. Kargas, N. D. Sidiropoulos, and X. Fu, "Tensors, Learning, and "Kolmogorov Extension" for Finite-Alphabet Random Vectors," *IEEE Trans. Signal Process.*, vol. 66, no. 18, pp. 4854–4868, Sep. 2018.

Marginal-Based Estimation: LS Criterion

- Estimate the third-order marginal PMF empirically, i.e., by counting all co-occurrences of triplet combinations $(j,k,\ell) \in \{1,\ldots,N\}$
- Marginal PMF estimate $\widehat{\boldsymbol{\mathcal{X}}}_{jk\ell} \in \mathbb{R}^{I_j \times I_k \times I_\ell}$ can be approximated by

$$\widehat{\boldsymbol{\mathcal{X}}}_{jk\ell} \approx [\![\boldsymbol{\lambda}; \boldsymbol{A}_j, \boldsymbol{A}_k, \boldsymbol{A}_\ell]\!] \tag{9}$$

- Full joint PMF obtained by solving a least-squares (LS) coupled tensor factorization problem using AO-ADMM [1]

$$\begin{aligned} \min_{\{\boldsymbol{A}_n\}_{n=1}^N, \boldsymbol{\lambda}} \quad & \sum_j \sum_{k>j} \sum_{\ell>k} \frac{1}{2} \left\| \underbrace{\widehat{\boldsymbol{\mathcal{X}}}_{jk\ell}}_{\text{Third-order sub-tensor (empirical estimate)}} - \underbrace{[\![\boldsymbol{\lambda}; \boldsymbol{A}_j, \boldsymbol{A}_k, \boldsymbol{A}_\ell]\!]}_{\text{Third-order CPD (to fit to empirical estimate)}} \right\|_{\mathsf{F}}^2 \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq \boldsymbol{0}, \ \boldsymbol{1}^{\mathsf{T}} \boldsymbol{\lambda} = 1 \\ & \boldsymbol{A}_n \geq \boldsymbol{0}, \ n = 1, \ldots, N \\ & \boldsymbol{1}^{\mathsf{T}} \boldsymbol{A}_n = \boldsymbol{1}^{\mathsf{T}}, \ n = 1, \ldots, N \end{aligned} \tag{10}$$

[1] N. Kargas, N. D. Sidiropoulos, and X. Fu, “Tensors, Learning, and “Kolmogorov Extension” for Finite-Alphabet Random Vectors,” *IEEE Trans. Signal Process.*, vol. 66, no. 18, pp. 4854–4868, Sep. 2018.

Marginal-Based Estimation

- The **LS criterion** [1]
  - ⇒ offers relatively convenient optimization procedures
  - ⇒ does not provide proper ‘weighting’ – may attribute zero probability to an element of the estimated PMF even when the empirical probability is non-zero
- The **Kullback-Leibler divergence (KLD)** is a more suitable criterion in the context of PMF estimation [2]
  - ⇒ treats smaller empirical probabilities in a different way than larger ones
  - ⇒ forbids a solution that attributes zero probability to an element of the estimated PMF where the empirical probability is non-zero

[2] A. Yeredor and M. Haardt, “Estimation of a Low-Rank Probability-Tensor from Sample Sub-Tensors via Joint Factorization Minimizing the Kullback-Leibler Divergence,” in *Proc. 27th European Signal Processing Conference (EUSIPCO)*, A Coruna, Spain, Sep. 2019.

Marginal-Based Estimation: KLD Criterion

- The KLD between two PMFs $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ and $\mathcal{Y} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ is

$$\begin{aligned} D(\mathcal{X} \| \mathcal{Y}) &= -\sum_{i_1,i_2,i_3} \mathcal{X}(i_1,i_2,i_3) \log \frac{\mathcal{Y}(i_1,i_2,i_3)}{\mathcal{X}(i_1,i_2,i_3)} \geq 0 \\ &= \sum_{i_1,i_2,i_3} \left(\mathcal{X}(i_1,i_2,i_3) \log \mathcal{X}(i_1,i_2,i_3) - \mathcal{X}(i_1,i_2,i_3) \log \mathcal{Y}(i_1,i_2,i_3) \right), \end{aligned} \tag{11}$$

  with equality if and only if $\mathcal{X} = \mathcal{Y}$

- If $\mathcal{X}$ is fixed, minimizing (11) w.r.t. $\mathcal{Y}$ amounts to minimizing

$$\begin{aligned} D(\mathcal{X} \| \mathcal{Y}) &= C - \sum_{i_1,i_2,i_3} \mathcal{X}(i_1,i_2,i_3) \log \mathcal{Y}(i_1,i_2,i_3) \\ &= C + \| \mathcal{X} \odot \log \mathcal{Y} \|_1 \end{aligned} \tag{12}$$

[2] A. Yeredor and M. Haardt, "Estimation of a Low-Rank Probability-Tensor from Sample Sub-Tensors via Joint Factorization Minimizing the Kullback-Leibler Divergence," in *Proc. 27th European Signal Processing Conference (EUSIPCO)*, A Coruna, Spain, Sep. 2019.

Marginal-Based Estimation: KLD Criterion

- Thus, the following KLD-based coupled tensor factorization problem is posed

$$\begin{aligned} \min_{\{\boldsymbol{A}_n\}_{n=1}^N, \boldsymbol{\lambda}} \quad & \sum_j \sum_{k>j} \sum_{\ell>k} \left\| \underbrace{\widehat{\mathcal{X}}_{jk\ell}}_{\substack{\text{Third-order sub-tensor} \\ \text{(empirical estimate)}}} \odot \log \underbrace{\left([\![\boldsymbol{\lambda}; \boldsymbol{A}_j, \boldsymbol{A}_k, \boldsymbol{A}_\ell]\!] \right)}_{\substack{\text{Third-order CPD} \\ \text{(to fit to empirical estimate)}}} \right\|_1 \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq \boldsymbol{0}, \ \boldsymbol{1}^\mathsf{T} \boldsymbol{\lambda} = 1 \\ & \boldsymbol{A}_n \geq \boldsymbol{0}, \ n = 1, \ldots, N \\ & \boldsymbol{1}^\mathsf{T} \boldsymbol{A}_n = \boldsymbol{1}^\mathsf{T}, \ n = 1, \ldots, N \end{aligned} \tag{13}$$

- Non-convex problem rewritten as two convex subproblems and solved using reparameterization techniques [2]

- The KLD criterion outperforms the LS criterion with respect to estimating the true PMF.

[2] A. Yeredor and M. Haardt, "Estimation of a Low-Rank Probability-Tensor from Sample Sub-Tensors via Joint Factorization Minimizing the Kullback-Leibler Divergence," in *Proc. 27th European Signal Processing Conference (EUSIPCO)*, A Coruna, Spain, Sep. 2019.

Outline

- Motivation for Estimating the joint PMF Tensor of a Set of Random Variables
- Modeling of the Naïve Bayes Model via the CPD
- Joint PMF Tensor Estimation Methods
  - ⇒ Marginal-based Estimation
    - LS (Least Squares) Criterion
    - KLD (Kullback-Leibler Divergence) Criterion
  - ⇒ Data-based Estimation
    - AD (Alternating Directions) Minimization
    - EM (Expectation Maximization)
- Simulation Results
  - ⇒ Synthetic Data Experiments
    - Speeding up the Convergence of EM: SQUAREM
  - ⇒ Real Data Experiments
- Conclusions

Data-Based Estimation

- Using the observed data, the maximum-likelihood (ML) estimate of the joint PMF tensor can be obtained [3]

  ⇒ more efficient – no marginals required

  ⇒ In the large-sample limit, maximizing likelihood corresponds to the minimization of the KLD between the observed data and the parametrized model distribution

[3] A. Yeredor and M. Haardt, “Maximum Likelihood Estimation of a Low-Rank Probability Mass Tensor From Partial Observations,” *IEEE Signal Process. Lett.*, vol. 26, no. 10, pp. 1551–1555, Oct. 2019.

Ilmenau University of Technology
Communications Research Laboratory

Data-Based Estimation: Problem Formulation

- Consider a discrete random vector $\boldsymbol{X} = [X_1, \ldots, X_N]^\mathsf{T} \in \mathbb{R}^N$ where each $X_n \in [1, I_n],\ n = 1, \ldots, N$
- Let $\boldsymbol{Y} = [Y_1, \ldots, Y_N]^\mathsf{T} \in \mathbb{R}^N$ be a random vector obtained as a partial observation of $\boldsymbol{X}$ such that (independently for each n)

$$Y_n = \begin{cases} X_n & \text{with probability } 1-p \\ 0 & \text{with probability } p \end{cases}, \quad n = 1, \ldots, N$$

$p \rightarrow$ outage probability

- **Goal**: estimate joint PMF tensor $\boldsymbol{\mathcal{X}}$ from T i.i.d. observations of $\boldsymbol{Y}$

[3] A. Yeredor and M. Haardt, "Maximum Likelihood Estimation of a Low-Rank Probability Mass Tensor From Partial Observations," *IEEE Signal Process. Lett.*, vol. 26, no. 10, pp. 1551–1555, Oct. 2019.

Data-Based Estimation

- Let $\boldsymbol{y}$ be a realization of $\boldsymbol{Y}$
- If $B \in [0, N]$ is the number of non-zero elements in $\boldsymbol{y}$ with indices $n_1, \ldots, n_B$, then,

$$\begin{aligned} \Pr(\boldsymbol{y}) &= p^{N-B}(1-p)^B \Pr(X_{n_1} = y_{n_1}, \ldots, X_{n_B} = y_{n_B}) \\ &= q \sum_{r=1}^{R} \boldsymbol{\lambda}(r) \boldsymbol{A}_{n_1}(y_{n_1}, r) \cdots \boldsymbol{A}_{n_B}(y_{n_B}, r) \\ &= q \sum_{r=1}^{R} \boldsymbol{\lambda}(r) \prod_{b=1}^{B} \boldsymbol{A}_{n_b}(y_{n_b}, r) \end{aligned} \tag{15}$$

- Factors corresponding to zero (unobserved) elements are marginalized out

[3] A. Yeredor and M. Haardt, "Maximum Likelihood Estimation of a Low-Rank Probability Mass Tensor From Partial Observations," *IEEE Signal Process. Lett.*, vol. 26, no. 10, pp. 1551–1555, Oct. 2019.

Data-Based Estimation

- Example: $N = 4$ and $\boldsymbol{y} = [3\ 4\ 0\ 2]^{\mathsf{T}} \Longrightarrow B = 3$

$$
\begin{aligned}
\Pr(\boldsymbol{y}) &= p(1-p)^3 \Pr(X_1 = 3, X_2 = 4, X_4 = 2) \\
&= q \sum_{i_3=1}^{I_3} \boldsymbol{\mathcal{X}}(3, 4, i_3, 2) \\
&= q \sum_{i_3=1}^{I_3} \sum_{r=1}^{R} \boldsymbol{\lambda}(r) \boldsymbol{A}_1(3, r) \boldsymbol{A}_2(4, r) \boldsymbol{A}_3(i_3, r) \boldsymbol{A}_4(2, r) \\
&= q \sum_{r=1}^{R} \boldsymbol{\lambda}(r) \boldsymbol{A}_1(3, r) \boldsymbol{A}_2(4, r) \underbrace{\left(\sum_{i_3=1}^{I_3} \boldsymbol{A}_3(i_3, r) \right)}_{=1} \boldsymbol{A}_4(2, r) \\
&= q \sum_{r=1}^{R} \boldsymbol{\lambda}(r) \boldsymbol{A}_1(3, r) \boldsymbol{A}_2(4, r) \boldsymbol{A}_4(2, r) \qquad (16)
\end{aligned}
$$

[3] A. Yeredor and M. Haardt, "Maximum Likelihood Estimation of a Low-Rank Probability Mass Tensor From Partial Observations," *IEEE Signal Process. Lett.*, vol. 26, no. 10, pp. 1551–1555, Oct. 2019.

Data-Based Estimation

- The log-likelihood of $\boldsymbol{y}$ is given by

$$\log \Pr(\boldsymbol{y}) = C + \log \sum_{r=1}^{R} \boldsymbol{\lambda}(r) \prod_{b=1}^{B} \boldsymbol{A}_{n_b}(y_{n_b}, r) \qquad (17)$$

where $C = \log q$

- For T i.i.d. observations $\boldsymbol{y}[1], \ldots, \boldsymbol{y}[T]$, the joint log-likelihood is (ignore constant)

$$\log \Pr(\boldsymbol{y}[1], \ldots, \boldsymbol{y}[T]) = \sum_{t=1}^{T} \log \sum_{r=1}^{R} \boldsymbol{\lambda}(r) \prod_{b=1}^{B[t]} \boldsymbol{A}_{n_b[t]}(y_{n_b[t]}[t], r) \qquad (18)$$

- Optimization problem for ML estimation of $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_N, \boldsymbol{\lambda}$ is

$$
\begin{aligned}
\min_{\{\boldsymbol{A}_n\}_{n=1}^{N}, \boldsymbol{\lambda}} \quad & -\sum_{t=1}^{T} \log \left(\sum_{r=1}^{R} \boldsymbol{\lambda}(r) \prod_{b=1}^{B[t]} \boldsymbol{A}_{n_b[t]}(y_{n_b[t]}[t], r) \right) \\
\text{subject to} \quad & \boldsymbol{\lambda} > \boldsymbol{0},\ \boldsymbol{1}^{\mathsf{T}} \boldsymbol{\lambda} = 1 \\
& \boldsymbol{A}_n \geq \boldsymbol{0},\ n = 1, \ldots, N \\
& \boldsymbol{1}^{\mathsf{T}} \boldsymbol{A}_n = \boldsymbol{1}^{\mathsf{T}},\ n = 1, \ldots, N \qquad (19)
\end{aligned}
$$

[3] A. Yeredor and M. Haardt, "Maximum Likelihood Estimation of a Low-Rank Probability Mass Tensor From Partial Observations," *IEEE Signal Process. Lett.*, vol. 26, no. 10, pp. 1551–1555, Oct. 2019.

Data-Based Estimation

- KLD-based **alternating directions (AD)** minimization [2]

  ⇒ rewrite the non-convex problem as two convex subproblems, i.e., minimize w.r.t. $\boldsymbol{\lambda}$ and to each $\boldsymbol{A}_n$ separately and solve these tasks using reparametrization techniques

[2] A. Yeredor and M. Haardt, "Estimation of a Low-Rank Probability-Tensor from Sample Sub-Tensors via Joint Factorization Minimizing the Kullback-Leibler Divergence," in *Proc. 27th European Signal Processing Conference (EUSIPCO)*, A Coruna, Spain, Sep. 2019.

[3] A. Yeredor and M. Haardt, "Maximum Likelihood Estimation of a Low-Rank Probability Mass Tensor From Partial Observations," *IEEE Signal Process. Lett.*, vol. 26, no. 10, pp. 1551–1555, Oct. 2019.

Data-Based Estimation

- **Expectation-maximization (EM) [3]**
  - ⇒ arises due to the naïve Bayes model implied by the low-rank assumption
  - ⇒ define "complete data" $\boldsymbol{z}[t] = [\boldsymbol{y}^{\mathsf{T}}[t] \; s[t]]^{\mathsf{T}}$ where $s[t] \in \{1, \ldots, R\}$ is the unknown latent state
  - ⇒ E-step: denote $\boldsymbol{\lambda}$ and all $\boldsymbol{A}_n$ at the first iteration as $\boldsymbol{\theta}'$ and compute

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}') &= \mathbb{E}\{\boldsymbol{z}[1], \ldots, \boldsymbol{z}[T] | \boldsymbol{y}[1], \ldots, \boldsymbol{y}[T]; \boldsymbol{\theta}'\} \\ &= \sum_{r=1}^{R} \sum_{t=1}^{T} c_{t,r}(\boldsymbol{\theta}') \cdot \left(\log \boldsymbol{\lambda}(s[t]) + \sum_{b=1}^{B[t]} \log \boldsymbol{A}_{n_b[t]}(y_{n_b[t]}[t], s[t]) \right) \\ &= \sum_{r=1}^{R} \left(C'_r \log \boldsymbol{\lambda}(r) + \sum_{n=1}^{N} \sum_{i_n=1}^{I_n} K'_r(n, i_n) \log \boldsymbol{A}_n(i_n, r) \right) \qquad (22) \end{aligned}$$

$$c_{t,r}(\boldsymbol{\theta}') \triangleq \mathsf{Pr}(s[t] = r | \boldsymbol{y}[t]; \boldsymbol{\theta}'), \quad C'_r \triangleq \sum_{t=1}^{T} c_{t,r}(\boldsymbol{\theta}'), \quad K'_r(n, i_n) \triangleq \sum_{t: y_n[t] = i_n} c_{t,r}(\boldsymbol{\theta}'), \; \forall n = 1, \ldots, N, \; i_n = 1, \ldots, I_n$$

  - ⇒ M-step:

$$\boldsymbol{\theta}' \leftarrow \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}')$$

$$\Longrightarrow \boldsymbol{\lambda}(r) = \frac{C'_r}{\sum_{g=1}^{R} C'_g} \quad \text{and} \quad \boldsymbol{A}_n(i_n, r) = \frac{K'_r(n, i_n)}{\sum_{j=1}^{I_n} K'_r(n, j)} \qquad (23)$$

Data-Based Estimation

- **AD**
  - ⇒ optimal solution to inner convex problems achievable
  - ⇒ is computationally expensive
- **EM**
  - ⇒ computationally cheaper (closed-form updates)
  - ⇒ may converge slowly
- A hybrid AD-EM algorithm is adopted [3]
  - ⇒ run a few (expensive) AD iterations to get to the vicinity of the global optimum
  - ⇒ switch to computationally cheaper EM for refinement
- The hybrid AD-EM algorithm obtains more accurate PMF estimates than the AO-ADMM algorithm based on triplets and the LS criterion

Simulation Results

- More insight into hybrid AD-EM algorithm needed
  - ⇒ when exactly do we switch from AD to EM?
  - ⇒ is it sufficient to run EM alone?
  - ⇒ how can we accelerate EM convergence?
  - ⇒ how do the algorithms perform on real data?
- Extensive simulations carried out in two approaches
  - ⇒ synthetic-data experiments
  - ⇒ real-data experiments – movie recommendation

Outline

- Motivation for Estimating the joint PMF Tensor of a Set of Random Variables
- Modeling of the Naïve Bayes Model via the CPD
- Joint PMF Tensor Estimation Methods
  - ⇒ Marginal-based Estimation
    - LS (Least Squares) Criterion
    - KLD (Kullback-Leibler Divergence) Criterion
  - ⇒ Data-based Estimation
    - AD (Alternating Directions) Minimization
    - EM (Expectation Maximization)
- Simulation Results
  - ⇒ Synthetic Data Experiments
    - Speeding up the Convergence of EM: SQUAREM
  - ⇒ Real Data Experiments
- Conclusions

Synthetic-Data Experiments

- Comparison of performance for the LS criterion (Triplets (LS)) and the KLD criterion (AD-EM)

- Comparison of AD-EM and EM performance
  - ⇒ convergence of the log likelihood
  - ⇒ different parameter settings
    - number of observations T
    - outage probability p
  - ⇒ different initialization strategies

Experimental Setup

- $N = 5$ discrete random variables
- Alphabet size $I_n = 10, \ \forall n = 1, \ldots, N \implies \boldsymbol{\mathcal{X}} \in \mathbb{R}^{10\times 10\times 10\times 10\times 10}$
- Ground-truth CPD components $\{\boldsymbol{A}_n\}_{n=1}^{N} \in U(0,1), \ \boldsymbol{\lambda} \in U(0,1)$ and normalized
- Observed data $\boldsymbol{Y} \in \mathbb{R}^{T\times N}$ with outage probability p
- $\boldsymbol{A}_n$ and $\boldsymbol{\lambda}$ initialized randomly and normalized
- Results averaged over $L = 100$ independent trials

Performance Metrics

- Mean Squared Relative Error (MSRE)

$$\mathrm{MSRE} = \frac{1}{L}\sum_{\ell=1}^{L} \frac{\|\boldsymbol{\mathcal{X}}_\ell - \widehat{\boldsymbol{\mathcal{X}}}_\ell\|_\mathsf{F}^2}{\|\boldsymbol{\mathcal{X}}_\ell\|_\mathsf{F}^2} \tag{24}$$

- Kullback-Leibler Divergence (KLD)

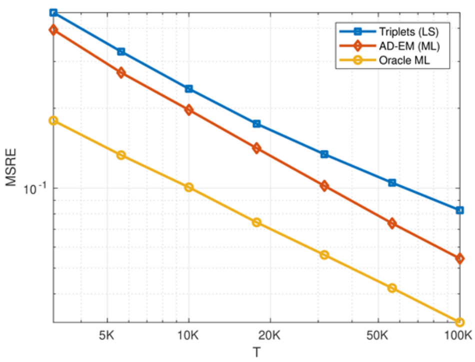
$$\mathrm{KLD} = \left\| \boldsymbol{\mathcal{X}} \odot \left(\log \widehat{\boldsymbol{\mathcal{X}}} - \log \boldsymbol{\mathcal{X}}\right) \right\|_1 \tag{25}$$

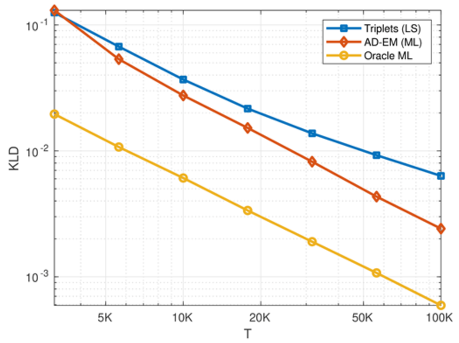
- Factor Match Score (FMS) [4]

$$\mathrm{FMS} = \max_{\sigma} \frac{1}{R} \sum_{r=1}^{R} \left(1 - \frac{|\boldsymbol{\lambda}_r - \hat{\boldsymbol{\lambda}}_{\sigma(r)}|}{\max\{\boldsymbol{\lambda}_r, \hat{\boldsymbol{\lambda}}_{\sigma(r)}\}}\right) \prod_{n=1}^{N} |\boldsymbol{a}_r^{(n)\mathsf{T}} \hat{\boldsymbol{a}}_{\sigma(r)}^{(n)}| \tag{26}$$

[4] E. C. Chi and T. G. Kolda, "On Tensors, Sparsity, and Nonnegative Factorizations," *SIAM J. Matrix Anal. & Appl.*, vol. 33, no. 4, pp. 1272–1299, Jan. 2012.

LS criterion vs. KLD criterion

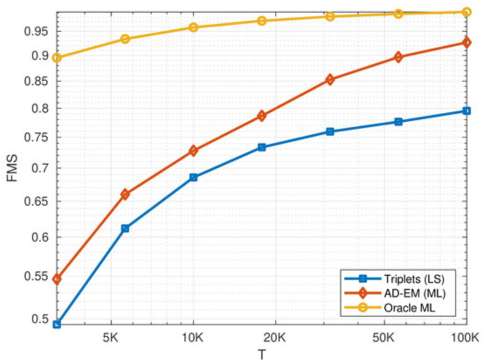




- Oracle ML – assumes the true latent state is known (provides lower bound)
- AD-EM (KLD criterion) outperforms the LS-based triplets (marginal) approach

[3] A. Yeredor and M. Haardt, "Maximum Likelihood Estimation of a Low-Rank Probability Mass Tensor From Partial Observations," *IEEE Signal Process. Lett.*, vol. 26, no. 10, pp. 1551–1555, Oct. 2019.

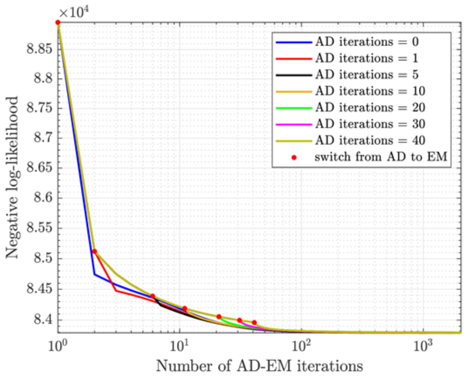
LS criterion vs. KLD criterion

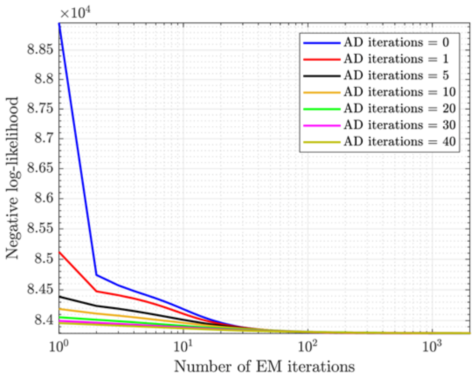


- Oracle ML – assumes the true latent state is known (provides lower bound)
- AD-EM (KLD criterion) outperforms the LS-based triplets (marginal) approach

[3] A. Yeredor and M. Haardt, "Maximum Likelihood Estimation of a Low-Rank Probability Mass Tensor From Partial Observations," *IEEE Signal Process. Lett.*, vol. 26, no. 10, pp. 1551–1555, Oct. 2019.

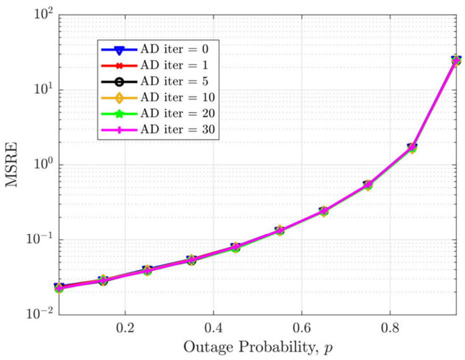
AD-EM vs. EM: Convergence

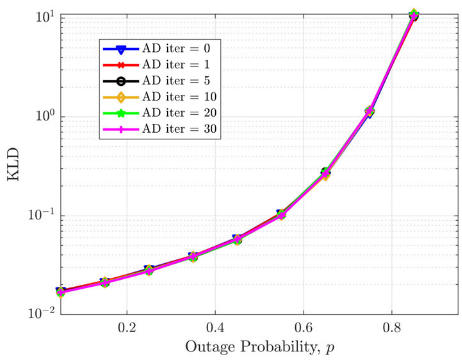




- Parameters: $T = 10^4,\ p = 0.25,\ R = 5$
- Convergence achieved with EM alone

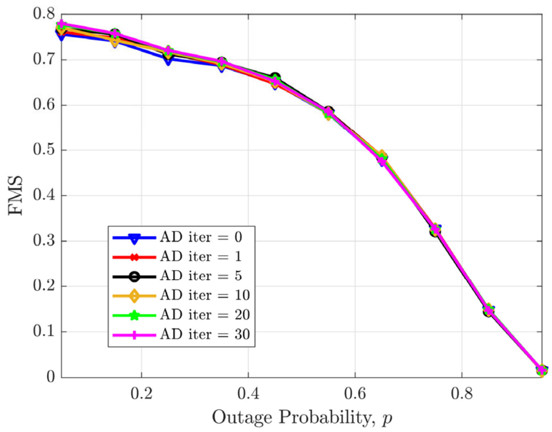
AD-EM vs. EM: Outage Probability





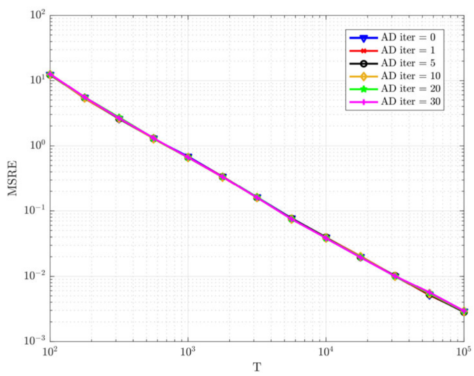
- Parameters: $T = 10^4,\ R = 5$
- Running EM alone is sufficient

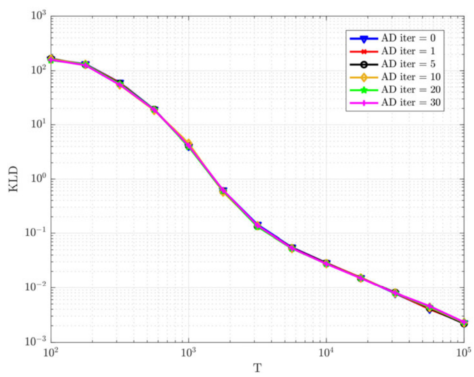
AD-EM vs. EM: Outage Probability



- Parameters: $T = 10^4, \;\; R = 5$
- Running EM alone is sufficient

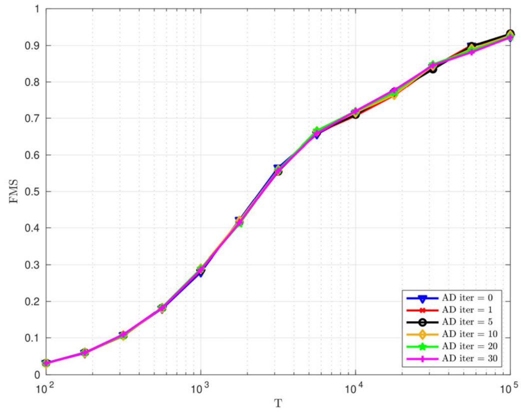
AD-EM vs. EM: Number of Observations





- Parameters: $p = 0.25, \;\; R = 5$
- Running EM alone is sufficient

AD-EM vs. EM: Number of Observations



- Parameters: $p = 0.25, \;\; R = 5$
- Running EM alone is sufficient

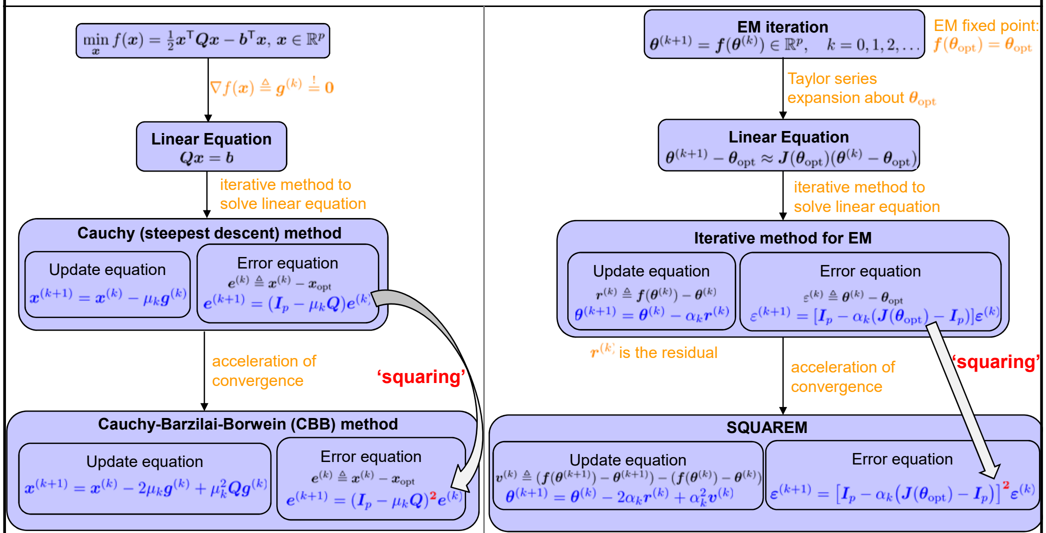
Outline

- Motivation for Estimating the joint PMF Tensor of a Set of Random Variables
- Modeling of the Naïve Bayes Model via the CPD
- Joint PMF Tensor Estimation Methods
  - ⇒ Marginal-based Estimation
    - LS (Least Squares) Criterion
    - KLD (Kullback-Leibler Divergence) Criterion
  - ⇒ Data-based Estimation
    - AD (Alternating Directions) Minimization
    - EM (Expectation Maximization)
- Simulation Results
  - ⇒ Synthetic Data Experiments
    - Speeding up the Convergence of EM: SQUAREM
  - ⇒ Real Data Experiments
- Conclusions

- EM exhibits linear convergence
- Convergence rate decreases with an increase in the amount of missing data
- SQUAREM [5] applies a squaring technique [6] used to accelerate the convergence of the Cauchy (steepest descent) method for solving a system of linear equations
- Key ingredient: EM iteration steps (fixed-point mapping) can be approximated by a linear equation
- Thus, the squaring technique can also be applied to EM to speed up convergence
- We adapt SQUAREM to the context of PMF estimation → SQUAREM-PMF

[5] R. Varadhan and C. Roland, "Simple and Globally Convergent Methods for Accelerating the Convergence of Any EM Algorithm," *Scandinavian Journal of Statistics*, vol. 35, no. 2, pp. 335–353, 2008.

[6] M. Raydan, "Relaxed Steepest Descent and Cauchy-Barzilai-Borwein Method," *Optimization*, pp. 155–167, 2002.



Adapting SQUAREM: SQUAREM-PMF

- To preserve EM stability, the SQUAREM algorithm in [5] modifies the step size α such that the negative log-likelihood always decreases
- For our PMF estimation problem, the SQUAREM update equation should additionally fulfill the probability simplex constraints
- We adapt the refinement procedure in [5] to refine the step size α such that
  - $\Rightarrow$ EM stability is preserved
  - $\Rightarrow$ probability simplex constraints are fulfilled

[5] R. Varadhan and C. Roland, “Simple and Globally Convergent Methods for Accelerating the Convergence of Any EM Algorithm,” *Scandinavian Journal of Statistics*, vol. 35, no. 2, pp. 335–353, 2008.

Adapting the SQUAREM step size α_k [7]

Adaptation of α_k

if $\alpha_k > -1$
    $\alpha_k = -1$
    $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}_0 - 2\alpha_k \boldsymbol{r}^{(k)} + \alpha_k^2 \boldsymbol{v}^{(k)} = \boldsymbol{\theta}_2$
else
    $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}_0 - 2\alpha_k \boldsymbol{r}^{(k)} + \alpha_k^2 \boldsymbol{v}^{(k)}$
    if nonnegativity constraint violated
        $\alpha_{k,i} = \frac{r_i^{(k)} \pm \sqrt{r_i^{(k)2} - v_i^{(k)} \theta_{0,i}}}{v_i^{(k)}}, \; i = 1, \ldots, p$
        find the legal segment(s) on the α_k-axis
        $\bar{\alpha}_k \leftarrow$ segment edge closest to optimal α_k } find step size that ensures nonnegativity
        if $\bar{\alpha}_k > \alpha_k$
            $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}_0 - 2\bar{\alpha}_k \boldsymbol{r}^{(k)} + \bar{\alpha}_k^2 \boldsymbol{v}^{(k)}$ } to avoid many likelihood computations later
        else
            $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}_0 - 2\alpha_k \boldsymbol{r}^{(k)} + \alpha_k^2 \boldsymbol{v}^{(k)}$
            $\boldsymbol{\theta}^{(k)} \leftarrow \mathcal{P}(\boldsymbol{\theta}^{(k)})$ } projection onto probability simplex, e.g., using [8]
        end if
    end if
    compute $L(\boldsymbol{\theta}^{(k)})$ } compute current negative log-likelihood
    while $L(\boldsymbol{\theta}^{(k)}) > L(\boldsymbol{\theta}^{(k-1)})$
        $\alpha_k = (\alpha_k - 1)/2$
        $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}_0 - 2\alpha_k \boldsymbol{r}^{(k)} + \alpha_k^2 \boldsymbol{v}^{(k)}$
        $\boldsymbol{\theta}^{(k)} \leftarrow \mathcal{P}(\boldsymbol{\theta}^{(k)})$ } projection onto probability simplex, e.g., using [8]
        compute $L(\boldsymbol{\theta}^{(k)})$
    end while
end if

[7] J. K. Chege, M. J. Grasis, A. Manina, A. Yeredor, and M. Haardt, “Efficient probability mass function estimation from partially observed data,” in *Proc. of 56th Asilomar Conference on Signals, Systems, and Computers,* Pacific Grove, CA, Nov. 2022.

[8] Y. Chen and X. Ye, “Projection Onto A Simplex,” arXiv, Feb. 09, 2011 [Online]. Available: http://arxiv.org/abs/1101.6081

The SQUAREM-PMF Algorithm [7]

SQUAREM-PMF

Input: $\{\boldsymbol{A}_{n,0}\}_{n=1}^{N}$, $\boldsymbol{\lambda}_0$

Output: ML estimates $\{\hat{\boldsymbol{A}}_n\}_{n=1}^{N}$, $\hat{\boldsymbol{\lambda}}$

Construct $\boldsymbol{\theta}_0 = [\mathrm{vec}^{\mathsf{T}}(\boldsymbol{A}_{1,0}), \ldots, \mathrm{vec}^{\mathsf{T}}(\boldsymbol{A}_{N,0}), \boldsymbol{\lambda}_0^{\mathsf{T}}]^{\mathsf{T}}$ } parameter vector

for $k = 1, \ldots, K$

$\quad \boldsymbol{\theta}_1 = \boldsymbol{f}(\boldsymbol{\theta}_0)$

$\quad \boldsymbol{\theta}_2 = \boldsymbol{f}(\boldsymbol{\theta}_1)$ } two EM updates

$\quad \boldsymbol{r}^{(k)} = \boldsymbol{\theta}_1 - \boldsymbol{\theta}_0$

$\quad \boldsymbol{v}^{(k)} = (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) - \boldsymbol{r}^{(k)}$ } two-point linear approx. of EM fixed point

$\quad \alpha_k = -\|\boldsymbol{r}^{(k)}\|/\|\boldsymbol{v}^{(k)}\|$

$\quad$ modify α_k using the proposed adaptation procedure } ensures stability is preserved and prob. simplex constraints are fulfilled

$\quad \boldsymbol{\theta}_0 = \boldsymbol{f}(\boldsymbol{\theta}^{(k)})$ } final EM update

$\quad$ **if** $\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}_0\| < \epsilon$

$\quad\quad$ Reconstruct $\{\boldsymbol{A}_n^{(k)}\}_{n=1}^{N}$ and $\boldsymbol{\lambda}^{(k)}$ from $\boldsymbol{\theta}^{(k)}$

$\quad\quad$ **break**

$\quad$ **end if**

end for

[7] J. K. Chege, M. J. Grasis, A. Manina, A. Yeredor, and M. Haardt, "Efficient probability mass function estimation from partially observed data," in *Proc. of 56th Asilomar Conference on Signals, Systems, and Computers,* Pacific Grove, CA, Nov. 2022.

SQUAREM-PMF Performance: Setup

- Max no. of EM iterations $\rightarrow K = 30000$
- Max no. of SQUAREM-PMF iterations $\rightarrow K = 10000$
- AD iterations = 20
- AO iterations = 10^3, ADMM iterations = 10^5
- Stopping criterion $\rightarrow \|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}\| < 10^{-7}$
- Parameters $\rightarrow \boldsymbol{\mathcal{X}} \in \mathbb{R}^{10\times10\times10\times10\times10},\ R = 5,\ p = 0.25,\ T = 10^5$
- Number of independent trials = 1000

SQUAREM-PMF Performance: Metrics

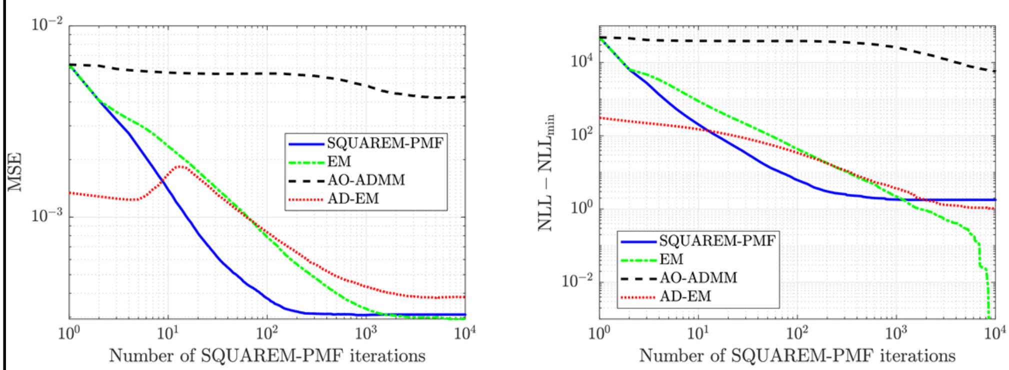
- Mean Squared Error (MSE) between the current estimate of the parameter vector $\boldsymbol{\theta}^{(k)} \in \mathbb{R}^M$ and the true parameter vector $\boldsymbol{\theta} \in \mathbb{R}^M$, averaged over L independent trials

$$\mathrm{MSE} = \frac{1}{L} \sum_{\ell=1}^{L} \frac{\|\boldsymbol{\theta}_\ell^{(k)} - \boldsymbol{\theta}\|^2}{M} \qquad (27)$$

- Negative Log-Likelihood (NLL) minus the overall minimum $\mathrm{NLL}_{\mathrm{min}}$ (across all four algorithms)

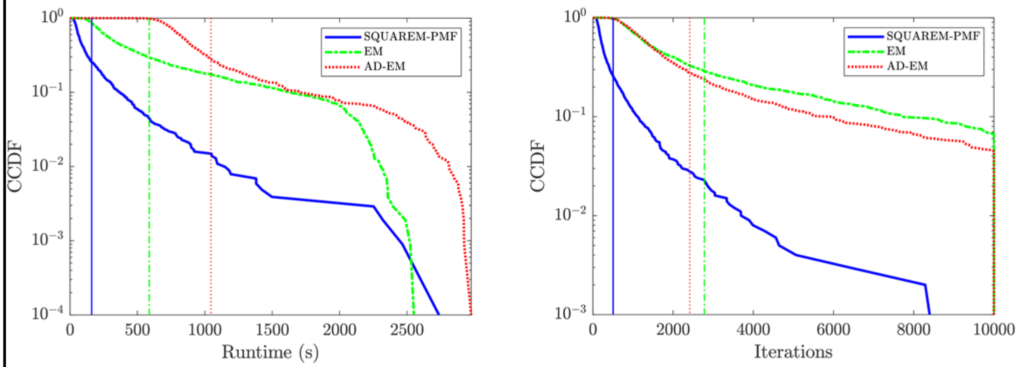
- Complementary Cumulative Distribution Function (CCDF) of the runtime and the number of iterations

SQUAREM-PMF Performance



- SQUAREM-PMF converges faster and preserves EM stability
- Initial AD iterations not required – EM alone is sufficient

SQUAREM-PMF Performance



	AD-EM	EM	SQUAREM-PMF
Average runtime [s]	1046	588	161
Average no. of iterations	2415	2779	505
Convergence success rate	954	932	1000

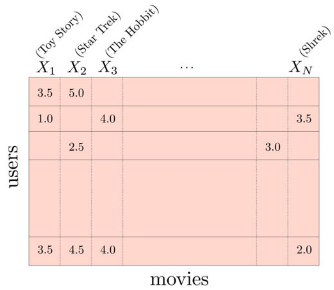
SQUAREM-PMF runs approx. 3.6 times faster than EM with far fewer iterations

Outline

- Motivation for Estimating the joint PMF Tensor of a Set of Random Variables
- Modeling of the Naïve Bayes Model via the CPD
- Joint PMF Tensor Estimation Methods
  - ⇒ Marginal-based Estimation
    - LS (Least Squares) Criterion
    - KLD (Kullback-Leibler Divergence) Criterion
  - ⇒ Data-based Estimation
    - AD (Alternating Directions) Minimization
    - EM (Expectation Maximization)
- Simulation Results
  - ⇒ Synthetic Data Experiments
    - Speeding up the Convergence of EM: SQUAREM
  - ⇒ Real Data Experiments
- Conclusions

Real-Data Experiments: MovieLens Dataset

- Context: movie recommendation using the 20M MovieLens dataset [9]
- Contains 20M movie ratings on a half-star scale {0.5,...,5.0}
- We select $N = 10$ top-rated movies from action, animation, and comedy genres
- Ratings mapped to the scale {1,...,10}



$\boldsymbol{Y} \in \mathbb{R}^{\text{users} \times \text{movies}}$

[9] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, Dec. 2015. [Online]. Available: https://doi.org/10.1145/2827872

Real-Data Experiments: Train/Val/Test Approach

- Initial dataset:
  - ⇒ 3 genres (action, animation, comedy)
  - ⇒ 10 top-rated movies
  - ⇒ users who have rated > 2 movies

Dataset	# users
Initial	84751
Training	59325
Validation	8476
Testing	16950

- For each trial ($L = 20$ independent trials):
  - ⇒ shuffle users in initial dataset
  - ⇒ split into train/val/test datasets (70 % / 10 % / 20 %)
  - ⇒ training – estimate $\{\boldsymbol{A}_n\}_{n=1}^N$ and $\boldsymbol{\lambda}$
  - ⇒ validation – find best parameters (rank, learning rate (AO-ADMM)) by hiding one rating and predicting it using various models
  - ⇒ retraining
    - use combined train/val dataset and best parameters
    - estimate $\{\boldsymbol{A}_n\}_{n=1}^N$ and $\boldsymbol{\lambda}$ again
  - ⇒ testing – hide one rating per user and predict it using the retrained model

Ilmenau University of Technology
Communications Research Laboratory

Real-Data Experiments: Prediction

- Ratings of t-th user : $\boldsymbol{y}[t] = [y_{1,t}, \ldots, y_{N,t}]$ and $y_{n,t} = 0$ if unobserved
- Conditional expectation (**MMSE** estimate)

$$\hat{y}_{n,t} = \sum_{i_n=1}^{I_n} i_n \Pr(i_n | y_{1,t}, \ldots, y_{n-1,t}, y_{n+1,t}, \ldots, y_{N,t}) \tag{28}$$

- **MAP** estimate

$$\hat{y}_{n,t} = \underset{i_n \in \{1,\ldots,I_n\}}{\arg\max} \Pr(i_n | y_{1,t}, \ldots, y_{n-1,t}, y_{n+1,t}, \ldots, y_{N,t}) \tag{29}$$

Real-Data Experiments: Metrics

- Root Mean Squared Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{L \cdot T} \sum_{\ell=1}^{L} \sum_{t=1}^{T} |\hat{y}_{n,t,\ell} - y_{n,t,\ell}|^2} \tag{30}$$

- Mean Absolute Error (MAE)

$$\mathrm{MAE} = \frac{1}{L \cdot T} \sum_{\ell=1}^{L} \sum_{t=1}^{T} |\hat{y}_{n,t,\ell} - y_{n,t,\ell}| \tag{31}$$

Real-Data Experiments: Test Results (± std. dev.) 61

Algorithm	Estimator	RMSE	MAE
AO-ADMM	MMSE	0.8452±0.004	0.6595±0.003
	MAP	0.9715±0.006	0.6558±0.004
EM	MMSE	**0.8192±0.004**	**0.6430±0.003**
	MAP	0.9731±0.006	0.6498±0.005
SQUAREM-PMF	MMSE	**0.8193±0.004**	**0.6429±0.003**
	MAP	0.9735±0.006	0.6496±0.004
Global average		0.9385±0.004	0.7305±0.003
User average		0.9399±0.005	0.7270±0.004
Movie average		0.9399±0.005	0.7270±0.004
Random guess		2.1077±0.010	1.7145±0.009

Conclusions 62

- Maximum likelihood estimation of a low-rank PMF tensor from partial observations
- The KLD criterion is superior to the LS criterion in the context of PMF estimation
- Extensive testing of AD-EM and EM algorithms using synthetic data
  - ⇒ running EM only is sufficient
  - ⇒ desirable – AD is computationally expensive
- Improved convergence rate of EM
  - ⇒ SQUAREM-PMF speeds up EM by approx. 3.6 times
  - ⇒ EM stability preserved by SQUAREM-PMF
- Testing of EM and SQUAREM-PMF on real data
  - ⇒ SQUAREM-PMF outperforms AO-ADMM
  - ⇒ SQUAREM-PMF performs comparably to EM (but converges faster)

[7] J. K. Chege, M. J. Grasis, A. Manina, A. Yeredor, and M. Haardt, "Efficient probability mass function estimation from partially observed data," in *Proc. of 56th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 2022.